Project Name

EDA and Statistical Analysis of FIFA Case Study

—

Overview

This Statistics and EDA project is designed to train and test you on basic Data Exploratory and Statistical techniques used in the industry today. Apart from bringing you to speed with basic descriptive and inferential methods, you will also deep dive into a dataset and perform thorough cleaning and analysis

in order to draw useful business insights from the data. This will expose you to what data scientists do most often–Exploratory Data Analysis.

Goals

1. Using the core statistical theoretical concepts and knowledge to solve real time problem statements.
2. Visualize a real time industry scenario where one can use these statistical concepts.
3. Detailed data analysis and number crunching using statistics
4. Exhaustive report building using EDA and visualization techniques to help the business take decisions using insights from the data

Specifications

Part -I is concept based and walks you through various concepts of descriptive statistics, probability distributions and inferential statistics including confidence intervals and hypothesis testing.

Part -II on the other hand is dataset based and explore various data cleaning options, data analysis options and using EDA to derive deep and meaningful insights for the business

PART-A (Concept Based)--25 points

The following are the BMI of 50 young adults

17.5 18.0 36.8 31.7 31.7

17.3 24.3 47.7 38.5 17.0

23.7 16.5 25.1 17.4 18.0

37.6 19.7 21.4 28.6 21.6

19.3 20.0 16.9 25.2 19.8

25.0 17.2 20.4 20.1 29.1

19.1 25.2 23.2 25.9 24.0

41.7 24.0 16.8 26.8 31.4

16.9 17.2 24.1 35.2 19.1

22.9 18.2 25.4 35.4 25.5

Use this data for answering following questions where relevant

Q1. Compute the mean, median and the mode of the data
Q2. Compute the range , variance and standard deviation of BMI
Q3. Find the mean deviation for the data . The mean deviation is defined as below.

$$Mean\ deviation = \frac{\sum|X - \bar{X}|}{n}$$

Q4. Calculate the Pearson coefficient of skewness and comment on the skewness of the data
[A measure to determine the skewness of a distribution is called the Pearson coefficient of skewness. The formula is

$$Skewness = \frac{3(\bar{X} - MD)}{s}$$

where MD is the median and s the standard deviation

The value of the coefficient if skewness usually ranges from –3 to 3. When the distribution is symmetric, the coefficient is zero; when the distribution is positively skewed , the coefficient is positive, and when the distribution is negatively skewed the coefficient is negative.]
Q5. Count the number of data values that fall within one standard deviation of the mean. Compare this with the answer from Chebyshev's Theorem.
Q6. Find the three quartiles and the interquartile range (IQR).
Q7. Are there any outliers in the data set ?
Q8. Draw a boxplot of the dataset to confirm .
Q9. Find the percentile rank of the datapoint 25.0.
Q10. What is the probability that a young adult has a BMI above 25.0?
Q11. Create a frequency distribution for the data and visualize it appropriately
Q12. Create a probability distribution of the data and visualize it appropriately.
Q13. What is the shape of the distribution of this dataset? Create an appropriate graph to determine that. Take 100 random samples with replacement from this dataset of size 5 each. Create a sampling distribution of the mean age of customers. Compare with other sampling distributions of sample size 10, 15, 20, 25, 30. State your observations. Does it corroborate the Central Limit Theorem?

Q14. Treat this dataset as a binomial distribution where *p* is the probability that a young adult has a BMI above 25.0. What is the probability that out of a random sample of 10 young adults exactly 6 are having BMI greater than 25.0?

Q15. A study claims that 40% of all young adults have BMI greater than 25.0. Using the Normal approximation of a Binomial distribution, find the probability that in a random sample of 100 young adults exactly 50 of them will have will have a BMI is greater than 25.0.

[Note that the normal distribution can be used to approximate a binomial distribution if np>=5 and nq>=5 with the following correction for continuity P(X=z) = P(z-0.5 < X < z+0.5)]

Q16. Compute a 95% Confidence Interval for the true BMI of the population of young adults using appropriate distribution.(State reasons as to why did you use a *z* or *t* distribution)

Q17. A data scientist wants to estimate with 95% confidence the proportion of young adults having BMI greater than 25.0. A recent study showed that 40% of all young adults have BMI greater than 25.0. The data scientist wants to be accurate within 2% of the true proportion. Find the minimum sample size necessary.

Q18. The same data scientist wants to estimate the true proportion of young adults having BMI greater than 25.0. She wants to be 90% confident and accurate within 5% of true proportion. Find the minimum sample size necessary.

Q19. A researcher claims that currently 55% of all young adults have BMI greater than 25 . Test his claim with an alpha =0.05 if out of a random sample of 30 CEOs only 20 are having BMI above 25

Q 20. A data scientist is researching the hypothesis that there is no difference between BMI of public vs private schools students. So he collects data from the two schools and finds that the proportion of public school students whose BMI is above 25.0 is 31.8 % vs Private school students is 38.7 %. Suppose the data scientist got these values after interviewing 500 students of each school.

a. What hypothesis would he use to compare the proportions of students having BMI greater than 25.0 among both the schools.
b. What are critical values to be used?
c. What statistical test will be used to compare these proportions ?

d. Complete the test and obtain the P-value.
e. Summarize his conclusion based on the P-value.

PART-B (Dataset Based)--25 points

EDA and STATS Mini Project-FIFA Case Study :

A new football club named ‘Brussels United FC’ has just been inaugurated. This club does not have a team yet. The team is looking to hire players for their roster. Management wants to make such decisions using a data-based approach. During a recent hiring drive, you were selected for the Data Science team as a Junior data scientist. Your team has been tasked with creating a report which recommends players for the main team. To start with, a total of 15 players are required. Player data for all teams have been acquired from FIFA. This data contains information about the players, the clubs they are currently playing for, and various performance measures. There is a limited budget for hiring players. The team needs 20 possible players to choose from. You have been requested to formulate a report in order to help the management make a decision regarding potential players.

Data:

The data contains details for over 25490 players playing in various football clubs in Europe. It contains information on age, skill rating, wages, player value, etc. The files provided are as follows: fifa.csv – data file. fifa_ variable_information.csv - information on individual variables.

Data Preprocessing:

1. Import the necessary libraries and read the data.
2. Drop any columns that you deem unnecessary for analysis.

Hint: At least keep the following columns
['ID','Name','Age','Nationality','Overall','Potential','Value','Wage','Joined','Preferred Foot', 'Contract Valid Until', 'Height', 'Weight', 'Penalties', 'Release Clause',’ International Reputation’, ‘Position’]

We encourage you to perform an analysis including other variables apart from the above variables.

3. The following columns need to be converted for further analysis:

Column	Details	Required output
'Value'	Amount with Euro symbol as prefix and suffix 'K' or 'M' indicating thousands and millions respectively.	Convert to Float after getting rid of currency symbol and suffix.
'Wage'	Amount with Euro symbol as prefix and suffix 'K' or 'M' indicating thousands and millions respectively.	Convert to Float after getting rid of currency symbol and suffix.
'Joined'	Year as a string, in some cases complete date as string	Convert to int with only year
'Contract Valid Until'	Date as a string	Convert to datetime type
'Height'	In inches with a quotation mark	Convert to Float with decimal points
'Weight'	Contains the suffix lbs	Remove the suffix and convert to float
'Release Clause'	Amount with Euro symbol as prefix and suffix 'K' or 'M' indicating thousands and millions respectively.	Convert to Float after getting rid of currency symbol and suffix.

(You might encounter Nan values in the above columns. Pandas treat Nan values as a float. Please keep that in mind when making the conversions.)

4. Check the duplicate records and do appropriate treatments.
5. Check the variation of the features.

If you are performing variance and standard deviation. Kindly explain why the variances of the variables are higher than the standard deviation.

Also, explain which one tells the exact variation of the features.

Based on this analysis decide which feature is not needed.

6. Check for missing values and do imputations where necessary.Note: Do the appropriate imputation based on the distribution.

Hints:

Few analyses are listed below to decide which imputation method to perform.

1. Skewness level verification
2. Kurtosis level measurement identification

Explain what the Skewness and Kurtosis depicted.

3. Distribution plots like KDE, Distribution plot, Box plot, etc.
7. find out If there is any player above 2.0*IQR – Q1 and below 2.0*IQR + Q3. What would happen if the data has above or below the mentioned values and display the names of the players?
8. Check the Categorical variables and Find if there is any data imbalance in any column also find if any cardinality problem exists.

9. Generate pair plots for the following variables:

Overall, Value, Wage, International Reputation, Height, Weight, Release Clause

10. Generate a table containing the top 20 players ranked by Overall score and whose contract expires in 2020.

1. What would the average wage for this set of players be?
2. What is the average age?
3. Is there a correlation between the Overall rating and Value for these players? If Yes what kind of relationship that the features have, also explain why this kind of relationship could happen.

11. Generate tables containing the top 5 players by Overall rating for each unique position.

1. Are there any players appearing at more than one table? Please point out such players.
2. What is the average wage one can expect to pay for the top 5 in every position?

Statistical Analysis:

1. Test statistically whether the Left-hand player's overall rating is higher than the Right-hand overall score. Alpha = 0.05

Before checking the test, Kindly make sure data is normally distributed.

2. Does the age factor affect the player's potential? Check the claim that the players who are greater than 35, their potential will be lesser than those whose age is less than 35. Alpha = 0.05
3. Use the statistical test to check the relationship between the Preferred Foot and Position with the 99% confident interval.
4. Does the International Reputation cause a significant effect on players' Wages? Check the claim with a 0.04 significance level. Check the Normality of data before the actual test.
5. Check the claim that the median wages of under top 20 players are lesser than or equal to 25000. Test the claim with a 0.05 % significance level. Check the data is normally distributed or not before the testing the claim statistically.